日常タスクにおける将来イベントのクロスモーダル説明文生成

Relational Future Captioning Model for Daily Tasks

神原 元就 *1　　杉浦 孔明 *1
Motonari Kambara　　Komei Sugiura

*1慶應義塾大学
Keio University

In this paper, our aim is to generate a caption about a future event. We propose the Relational Future Captioning Model (RFCM), a crossmodal language generation model for the future captioning task. The RFCM has the Relational Self-Attention Encoder to extract the relationships between events more effectively than the conventional self-attention in transformers. We conducted comparison experiments, and the results show the RFCM outperforms a baseline method on two datasets.

1. はじめに

日常タスクを支援するため，ユーザと自然にコミュニケーションができる生活支援ロボット (Domestic Srevice Robot, DSRs) の実用化は要支援者にとって有望な解決策の一つである．DSRs にとって，日常タスクの支援における安全性及び，利便性は重要である．日常タスクにおいて，動作実行前にタスクの実行に伴う危険性を予測し，ユーザに判断を仰ぐ能力はこの安全性，及び利便性を高める．例えば，物体を配置する際，他の物と衝突した場合，連鎖的に衝突が起こり物体が破損する危険性がある．そうした危険性について DSRs が事前に予測し，自然言語により注意喚起できることは重要である．一方で，この能力は未だに不十分である．

本論文では，時刻 t までのクリップを基に時刻 t+1 のイベントについての説明文を生成するタスクである，Future captioning タスク [Hosseinzadeh 21] を扱う．特に，日常生活における Future captioning タスクを扱う．ここで，DSR がペットボトルを棚に置く際に「ハンドがマグカップに接触することで，マグカップがその隣りにあるグラスに更に接触し，グラスが倒れる危険性があります」のような文を動作実行前にユーザに提示することが望ましい．

このタスクは，モデルは将来のクリップを利用することができないという点で難しい．そのため，過去イベントを用いた将来イベントの予測，及びキャプションの生成という 2 つの要素が求められる．実際，4 章で示すように，動画キャプション生成タスクのためのモデルを適用した場合，正解文と生成文の品質には大きな差が見られる．

これは，既存手法では，イベントの視覚特徴系列と文の関係のモデル化が不十分であるためである．それらの手法における transformer の自己注意では，イベントの即時的な視覚特徴と文の間の関係がモデル化されている。また，多くの動画キャプション生成モデルは、次時刻以降のキャプションを用いて次時刻のキャプションを行うモデルもあるため Future captioning タスクには不適切である．

Future captioning タスクが含まれるキャプション生成分野には多くの研究がある [Wang 18, Krishna 17, Lei 20, Kambara 21, Magassouba 21]．このタスクにおける代表的な手法として [Hosseinzadeh 21, Mori 21, Mahmud 21] が挙げられる．

連絡先: 神原元就，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，motonari.k714@keio.jp

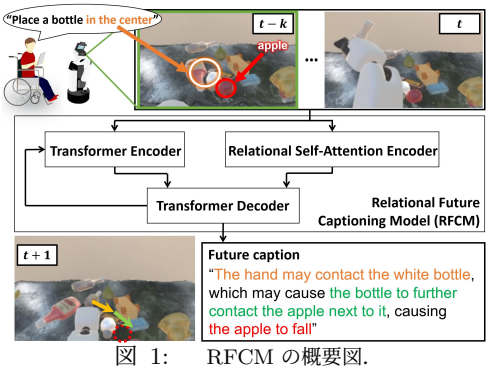

図 1:　RFCM の概要図.

提案手法は過去のイベント及び将来のイベントの関係性を適切に考慮したキャプションを生成できる．なぜなら，提案手法は，イベント間の関係性から適切な将来イベントのキャプションを生成するための source-target 注意機構を導入する．この機構では，過去のクリップに由来する特徴量を source，キャプション及びクリップに由来する特徴量を target とする．

提案手法は 3 つのモジュールから構成され，それぞれ Relational Self-Attention (RSA [Kim 21]) エンコーダ, transformer エンコーダ, および transformer デコーダである．提案手法は RSA エンコーダによって，過去のイベントとの関係性を適切に考慮した将来のイベント表現を獲得できる．なぜなら RSA は，既存の注意機構よりも効果的に，過去のイベントのうちどのイベントに注目すればよいかを学習できるためである．また，transformer デコーダは過去のイベントの関係性から，適切に将来イベントのキャプションを生成できる．transformer デコーダは，RSA エンコーダの出力をクエリとし，transformer エンコーダの出力をキー，及びバリューとする source-target 注意機構を持つ．これによって，過去のイベント間の関係性を自然言語に適切に接地できる．

本研究の独自性は以下である．

- Future captioning タスクのためのクロスモーダル言語生成モデル，RFCM を提案する．
- RSA エンコーダの導入により，既存の自己注意機構に比べ，より効果的にイベント間の関係性を抽出できる．

2. 問題設定

本論文では Future captioning タスクを扱う．Future captioning タスクでは時刻 t までのクリップから時刻 $t+1$ におけ

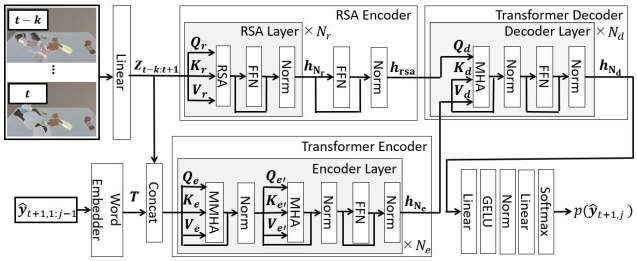

図 2: RFCM のネットワーク図.

るクリップの説明文を生成することを目的とする． 図 1 に本タスクの典型例を示す．本タスクでは，図 1 に示したように，時刻 t までのクリップから時刻 $t+1$ での状況を推論し説明文を出力することが望ましい． 本タスクにおいて，モデルへの入力は時刻 t までのクリップであり，出力は時刻 $t+1$ に起きる出来事についての説明文である．ここで，本稿ではクリップをイベントを描写したフレーム系列と定義する．本稿では，他タスクでの事前学習は許可されていないものとする．これは，本研究は知識転移を目的としたものではないためである．

3. 提案手法

ネットワークの構造を図 2 に示す．図において，Norm はレイヤ正規化 (layer normalization, LN) 層を表す．ネットワーク入力は以下で表される，時刻 t-k から時刻 t までの動画クリップとする．

ネットワーク入力 $\boldsymbol{x}$ は $\boldsymbol{x} = \{\boldsymbol{x}_{t-k}, ..., \boldsymbol{x}_t\}$, で定義される．ここで，$\boldsymbol{x}_t \in \mathbb{R}^{d_{\text{in}}}$ 及び $k \in \mathbb{N}$ はそれぞれ時刻 t におけるクリップ及び時刻 t 以前のクリップ数を表す．

$\boldsymbol{x}$ を基に，$\boldsymbol{Z}_{t-k:t+1} \in \mathbb{R}^{(k+2)\times d_{\text{in}}}$ を以下の式により獲得する．

$$\boldsymbol{z}_\tau = \begin{cases} f_z(\boldsymbol{x}_\tau) & (\tau > t) \\ f_z(\boldsymbol{x}_\tau, \boldsymbol{x}_t) & (\tau \leq t), \end{cases}$$

$$\boldsymbol{Z}_{t-k:t+1} = \{\boldsymbol{z}_{t-k}; ...; \boldsymbol{z}_{t+1}\},$$

ここで，$\tau = t-k, ..., t+1$, $f_z(\cdot)$ は線形変換を表す．$\boldsymbol{Z}_{t-k:t+1}$ はイベント系列に関する情報を含む.

3.1 RSA エンコーダ

RSA エンコーダは Relational Self-Attention [Kim 21] を利用した構造を持ち，イベント間の関係性を抽出する．RSA エンコーダは N_r 層の RSA 層から構成される．第一層目では，$\boldsymbol{Z}_{t-k:t+1}$ に対して [Vaswani 17] と同様の手順で，三角関数を用いた位置埋め込みを行う．

その後，クエリ $\boldsymbol{Q}_r \in \mathbb{R}^{d_{\text{rsa}}}$, キー $\boldsymbol{K}_r \in \mathbb{R}^{(k+2)\times d_{\text{rsa}}}$, 及びバリュー $\boldsymbol{V}_r \in \mathbb{R}^{(k+2)\times d_{\text{rsa}}}$ を以下のようにして獲得する．

$$\boldsymbol{Q}_r = \boldsymbol{z}_t,$$

$$\boldsymbol{K}_r = \boldsymbol{V}_r = \boldsymbol{Z}_{t-k:t+1},$$

ここで，d_{rsa} は各層の次元数を表す． 続いて，basic kernel $\boldsymbol{\varphi}_p \in \mathbb{R}^{k+2}$ 及び relational kernel $\boldsymbol{\varphi}_h \in \mathbb{R}^{k+2}$ を以下のようにして得る．

$$\boldsymbol{\varphi}_p = \boldsymbol{W}_p \boldsymbol{Q}_r,$$

$$\boldsymbol{Q}_{r'} = \{\boldsymbol{Q}_r; ...; \boldsymbol{Q}_r\} \in \mathbb{R}^{(k+2)\times d_{\text{rsa}}},$$

$$\boldsymbol{\varphi}_h = \boldsymbol{W}_h f_{\text{flatten}}(\boldsymbol{Q}_{r'} \odot \boldsymbol{K}_r),$$

ここで，$f_{\text{flatten}}(\cdot)$ はベクトル化を表す．続いて，relational context $\boldsymbol{\Phi}_g \in \mathbb{R}^{(k+2)\times d_{\text{rsa}}}$ を以下のようにして得る．

$$\boldsymbol{\Phi}_g = \boldsymbol{V}_r + \boldsymbol{W}_g \boldsymbol{V}_r^\top \boldsymbol{V}_r.$$

最終的に RSA $\boldsymbol{\varphi} \in \mathbb{R}^{d_{\text{rsa}}}$ を以下のようにして得る．

$$\boldsymbol{\varphi} = (\boldsymbol{\varphi}_p + \boldsymbol{\varphi}_h)^\top \boldsymbol{\Phi}_g.$$

RSA $\boldsymbol{\varphi}$ は入力イベント間の関係性についての情報を含む．$\boldsymbol{\varphi}$ を時刻 t における潜在特徴量として用いるため，$\boldsymbol{z}_t$ を $\boldsymbol{\varphi}$ によって置き換える． これにより，中間特徴量 $\boldsymbol{h}_r = \{\boldsymbol{z}_{t-k}; ...; \boldsymbol{z}_{t-1}; \boldsymbol{\varphi}; \boldsymbol{z}_{t+1}\} \in \mathbb{R}^{(k+2)\times d_{\text{rsa}}}$ を獲得する．各層の出力 $\boldsymbol{h}_{\text{n}_\text{r}}$ $(n_r = 1, ..., N_r)$ は $\boldsymbol{h}_r$ をフィードフォワードネットワーク (feedforward network, FFN) 及び LN 層を用いて変換することで得る．エンコーダの出力 $\boldsymbol{h}_{\text{rsa}} \in \mathbb{R}^{(k+2)\times d_{\text{rsa}}}$ は $\boldsymbol{h}_{rsa} = f_{\text{LN}}(f_{\text{FFN}}(\boldsymbol{h}_{\text{N}_\text{r}}))$ により得られる．ここで，$f_{\text{FFN}}(\cdot)$ 及び $f_{\text{LN}}(\cdot)$ はそれぞれ FFN 及び LN 層を表す.

3.2 Transformer エンコーダ／デコーダ

transformer エンコーダは N_e 層のエンコーダ層からなる．各層は Masked Multi-Head Attention (MMHA), Multi-Head Attention (MHA), 及び FFN 層から構成される．

エンコーダへの入力は $\boldsymbol{h}_c = \{\boldsymbol{Z}_{t-k:t+1}; \boldsymbol{T}\} \in \mathbb{R}^{(k+I+2)\times d_{\text{in}}}$ で表される．ここで，I は文の最大長を表す．訓練において，$\boldsymbol{T} \in \mathbb{R}^{I\times d_{\text{in}}}$ は時刻 $t+1$ における正解文, $\boldsymbol{y}_{t+1}$ を BERT 埋め込み器 [Devlin 19] により埋め込んだテキスト特徴量を表す．また，推論において j 番目の単語生成時, $\boldsymbol{T}$ は生成単語列 $\hat{\boldsymbol{y}}_{t+1,1:j-1}$ を埋め込んだテキスト特徴量を表す.

第 n_e+1 層への入力を $\boldsymbol{h}_{n_e}$ $(n_e = 0, ..., N_e - 1)$ と表す．また，$\boldsymbol{h}_0 = \boldsymbol{h}_c$ である．MMHA 層において，クエリ $\boldsymbol{Q}_e \in \mathbb{R}^{(k+I+2)\times d_e}$, キー $\boldsymbol{K}_e \in \mathbb{R}^{(k+I+2)\times d_e}$, 及びバリュー $\boldsymbol{V}_e \in \mathbb{R}^{(k+I+2)\times d_e}$ $(d_e = d_{\text{enc}}/N_h)$ はそれぞれ $\boldsymbol{Q}_e = \boldsymbol{W}_q^{(e)} \boldsymbol{h}_{n_e}$, $\boldsymbol{K}_e = \boldsymbol{W}_k^{(e)} \boldsymbol{h}_{n_e}$, 及び $\boldsymbol{V}_e = \boldsymbol{W}_v^{(e)} \boldsymbol{h}_{n_e} (e = 1, ..., N_h)$ のようにして得られる． ここで，N_h 及び d_{enc} はそれぞれアテンションヘッドの数及び各層の次元数を表す．訓練において, MMHA 層では, m 番目以降の単語トークンがマスクされる．これは，エンコーダが m 番目以降の単語トークンを用いて m 番目の単語トークンを生成することを防ぐためである (teacher forcing). 注意 $\boldsymbol{A}_{\text{MMHA}} \in \mathbb{R}^{(k+I+2)\times d_{\text{enc}}}$ を $\boldsymbol{Q}_e, \boldsymbol{K}_e$, and $\boldsymbol{V}_e$ を用いて獲得するため，[Vaswani 17] に示された計算を行う．

その後，MHA 層において, 注意 $\boldsymbol{A}_{\text{MHA}}$ が MMHA 層における計算と同様の計算で得られる．ここで，$\boldsymbol{h}_{\text{enc}}$, $\boldsymbol{Q}_e$, $\boldsymbol{K}_e$, 及び $\boldsymbol{V}_e$ はそれぞれ $\boldsymbol{A}_{\text{MMHA}}$, $\boldsymbol{Q}_{e'}$, $\boldsymbol{K}_{e'}$, 及び $\boldsymbol{V}_{e'}$ により置き換える．第 n_e+1 層のエンコーダ層の出力, $\boldsymbol{h}_{n_e+1} \in \mathbb{R}^{(k+I+2)\times d_{\text{enc}}}$, は $\boldsymbol{h}_{n_e+1} = f_{\text{LN}}(f_{\text{FFN}}(\boldsymbol{A}_{\text{MHA}}))$ により得られる．エンコーダの出力は $\boldsymbol{h}_{N_e}$ である.

transformer デコーダは $\boldsymbol{h}_{\text{rsa}}$ 及び $\boldsymbol{h}_{N_e}$ を入力とする．このモジュールは N_d 層のデコーダ層から構成される．各層の構造は transformer エンコーダにおける MHA 及び FFN 層から構成されたものである．一方で，クエリを $\boldsymbol{h}_{\text{rsa}}$ を基に作成し，キー及びバリューを $\boldsymbol{h}_{N_e}$ を基に作成するという違いがある．transformer デコーダからは $\boldsymbol{h}_{N_d} \in \mathbb{R}^{(k+I+2)\times d_{\text{dec}}}$ が得られる．ここで，d_{dec} は各層の次元数を示す.

最終的に，生成文における j 単語目の予測確率 $p(\hat{\boldsymbol{y}}_{t+1,j}) \in \mathbb{R}^{N_v}$ を $p(\hat{\boldsymbol{y}}_{t+1,j}) = \text{softmax}(f_{\text{gen}}(\boldsymbol{h}_{N_d}))$ により得る．ここで，N_v は辞書サイズを表す．また，$f_{\text{gen}}(\cdot)$ は全結合層，GELU 関数，LN 層，そして全結合層の順に適用する計算を表す.

損失関数 $\mathcal{L}$ は以下で定義される．

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{CE}}(y_{t+1}, p(\hat{\boldsymbol{y}}_{t+1})) + \lambda_{\text{iwp}}\mathcal{L}_{\text{iwp}}(\boldsymbol{y}_{t+1,1}, p(\hat{\boldsymbol{y}}_{t+1,1})) + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{mse}}\mathcal{L}_{\text{MSE}}(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}),$$

ここで，$\mathcal{L}_{\text{MSE}}(\cdot,\cdot)$ 及び $\mathcal{L}_{\text{CE}}(\cdot,\cdot)$ はそれぞれ平均二乗誤差関数及び交差エントロピー誤差関数を表す．また，λ_{ce}, λ_{corr}, λ_{mse} 及び λ_{iwp} はハイパーパラメータである．$\mathcal{L}_{\text{corr}}$ は $\hat{y}_{t+1}$ が時刻 t よりも早い，または遅い時刻のイベントについて述べている場合にペナルティを科す．$\mathcal{L}_{\text{iwp}}$ は生成文の 1 単語目，$y^*_{t+1,1}$ が適切でない場合にペナルティを科す． これは

$\mathcal{L}_{\mathrm{iwp}}(y_{t+1,1}, p(\hat{\boldsymbol{y}}_{t+1,1})) = \gamma_{\mathrm{iwp}}\mathcal{L}_{\mathrm{CE}}(y_{t+1,1}, p(\hat{\boldsymbol{y}}_{t+1,1}))$ のように定義される．ここで，$\gamma_{iwp} = 1/W$ である．

4. 実験

4.1 データセット

データセットとして，YouCook2-FC，及び BILA-caption データセットを利用した．

YouCook2-FC データセットは Future captioning タスクのためのデータセットの一つである．一般的に，料理において，次の調理手順はそれまでの手順により決定される．本研究では YouCook2 データセット [Zhou 18] に基づき YouCook2-FC データセットを構築した．訓練，検証，及びテスト集合のサイズはそれぞれ 7435, 1569, 3035 とした．

BILA-caption データセットは物体配置において起こりうる衝突についての Future captioning モデルを評価するため，新たに収集したデータセットである．図 1 に例を示す．データセットを構築するため，WRS2018 パートナーロボットチャレンジ／バーチャルスペースコンペティションにおいて使用された SIGVerse[Inamura+ 13] を拡張したものを使用した．シミュレーションでは，生活支援ロボットがランダムに選択されたボトルや缶などの日用品を 5 種類の机や棚などの家具の中央に配置する．生活支援ロボットのヘッドカメラから撮影した映像を収集した．それぞれのサンプルには “ the apple rolled over because the rabbit figure next to it was pushed by the robot ” のようなロボットによる配置の結果起きた状況を説明した文を付与した．データセットは 1000 本の動画および，衝突イベントに対して付与された英語の説明文 1000 文からなる．動画の合計時間および平均時間はそれぞれ 2.2 時間および 8 秒である．各動画について，アノテータによって文が一つずつ付与されている．語彙サイズは 245 語である．訓練，検証，及びテスト集合のサイズはそれぞれ 800, 100, 100 とした．

4.2 実験設定

以下にデータセットの事前処理における手順を示した．YouCook2-FC データセットにおいて，与えられた開始，及び終了時刻に基づきクリップを切り出した．BILA-caption データセットにおいて，開始時刻は DSR のアームが動き出した時刻とした．また，終了時刻は，開始時刻から 4 秒後，あるいは衝突イベントの発生のうちいずれかが満たされた時刻とした．

各クリップは，YouCook2-FC 及び BILA-caption データセットにおいて，それぞれ 0.6 及び 8fps に変換された．その後，YouCook2-FC データセットについては，[Ging 20] に示された手順により処理を行った．BILA-caption データセットについては，Howto100m データセット [Miech 19] で事前訓練済みの S3D [Miech 20] により 512 次元の特徴量とした後，全結合層によって 384 次元の特徴量とした．

実験設定は以下とした．最適化関数，学習率，バッチサイズ，及びエポック数はそれぞれ Adam (β_1: 0.9, β_2: 0.999), 1.0e-4, 16, 及び 25 とした．各モジュールにおいて，N_e, d_{enc}, N_h, N_d, d_{dec}, N_r, 及び d_{rsa} はそれぞれ，3, 384, 12, 3, 384, 2, 及び 384 とした． 損失関数において，λ_{ce}, λ_{iwp}, λ_{corr}，及び λ_{mse} はそれぞれ 30, 1.0, 0.1, 0.005, 及び 10 とした．$\mathcal{L}_{\mathrm{iwp}}$ において，n_{th} 回以上出現した単語を扱った．ここで，$n_{\mathrm{th}} = 30$ とした．また，YouCook2 及び BILA-caption データセットにおける W はそれぞれ，3000 及び 1000 とした．RFCM の訓練可能パラメータ数及び積和演算数はそれぞれ 3.1M 及び 540M であった．

訓練は 16GB メモリ搭載 NVIDIA Tesla V100 SXM2, 240 GB RAM, 及び Intel Xeon Gold 6148 プロセッサを用いて行われた．YouCook2-FC 及び BILA-caption データセットにおける訓練時間はそれぞれ 6.2 時間及び 1.6×10^{-1} 時間であった．同様に，推論時間はそれぞれ 1.7×10^{-2} 及び 4.9×10^{-2} s/サンプルであった．early stopping の条件として，[Prechelt 98] に示される generalization を用いた．

4.3 定量的結果

RFCM 及び Memory-Augmented Recurrent Transformer (MART [Lei 20]) について比較を行った．MART は transformer を用いた動画キャプション生成タスクにおける代表的な手法の一つであり，Future captioning タスクへの適用も可能であったことからベースライン手法とした．表 1 に YouCook2-FC 及び BILA-caption データセットにおける定量的結果を示した．各手法につき実験を 5 回行い，表にはその平均値及び標準偏差を示した．

生成文の評価は，動画キャプション生成タスクにおける標準尺度である BLEU4, ROUGE-L [Lin 04], METEOR [Banerjee 05], 及び CIDEr-D [Vedantam 15] により行った．主要尺度は CIDEr-D とした．

まず YouCook2-FC データセットにおいて比較した．表より，CIDEr-D が 0.46 ポイント向上した．また，BLEU4 及び METEOR もそれぞれ 0.18 ポイント及び 0.29 ポイント向上した．

続いて，BILA-caption データセットにおいて比較した．表より，CIDEr-D は 12.28 ポイントと大幅に向上した．また，全ての評価尺度においてスコアが向上した．これらの結果より，Future captioning タスクにおいて RFCM はベースライン手法よりも適切な文生成が可能であることがわかった．

4.4 定性的結果

図 3-4 に YouCook2-FC 及び BILA-caption データセットにおける定性的結果を示した．各図において，上段はイベントを示す．各イベントは時系列順に示されている．紙面の都合上いくつかのイベントについて省略した．また，下段は正解文及び各種法による生成文を示す．

図 3 に YouCook2-FC データセットにおける成功例を示した．図において，“rub flour” 及び “coat with breadcrumbs” という 2 つのイベントが起きた．ベースライン手法は “breadcrumbs” について言及しなかった．そのため，生成文は不足していた．一方で，提案手法は “coat the chicken with flour and bread crumbs” と適切に記述した．以上の結果から，提案手法は次時刻の調理ステップを適切に予測し，キャプションを生成できたといえる．

同様に，図 4 に BILA-caption データセットにおける成功例を示した．図において，把持されている物体は “the white bottle”，衝突する物体は “the camera” である．ベースライン手法は衝突する物体について “a black teapot” と誤って記述した． 一方で，提案手法ではそれぞれ “a white jar,” および “the camera” と適切に記述した．以上より，提案手法では動画に出現する物体についての特徴を適切に表現できることがわかった．

4.5 Ablation study

各モジュールについて ablation study を行った．表 1 に定量的結果を示した．どのモジュールが最も性能向上に寄与しているかを調査するため，ablation 条件を以下の 2 条件とした．(a) w/o RSA: RSA エンコーダにおいて，RSA 層の代わりに MHA 層 [Vaswani 17] を用いた．(b) w/o transformer decoder: transformer デコーダを取り除いた．

表 1: 各手法による生成文の，BLEU4, ROUGE-L, METEOR, CIDEr-D による評価結果

Methods	YouCook2-FC				BILA-caption			
	BLEU4↑	METEOR↑	ROUGE-L↑	CIDEr-D↑	BLEU4↑	METEOR↑	ROUGE-L↑	CIDEr-D↑
MART [Lei 20]	6.85±0.18	14.24±0.07	**30.80**±0.21	20.86±1.07	19.01±0.74	21.02±0.45	30.30±0.64	37.33±4.37
Ours (w/o RSA)	6.70±0.36	14.14±0.50	30.18±0.18	21.26±2.83	20.37±0.36	22.04±0.19	40.67±0.48	44.65±4.89
Ours (w/o Transformer Decoder)	6.68±0.13	14.09±0.15	30.16±0.31	19.75±1.41	21.08±1.62	22.39±0.84	40.92±1.32	45.05±6.72
Ours (RFCM)	**7.03**±0.15	**14.53**±0.09	30.49±0.21	**21.32**±1.09	**21.74**±1.02	**22.74**±0.57	**41.44**±0.86	**49.61**±8.02

Ref: “Place chicken on a plate or tray and season generously with mixed spices”

Baseline: “Cut the chicken into small pieces”

Ours: “Place the chicken pieces in a bowl”

図 3: YouCook2-FC データセットにおける成功例.

Ref: “Robot hits various things in the center because robot tried to put a white jar”

Baseline: “Robot hits the white bottle in the center hard because there was it in the robot’s orbit”

Ours: “The white jar is flipped because robot tried to put it on a teacup”

図 4: BILA-caption データセットにおける成功例.

YouCook2-FC データセットにおいて，提案手法と条件 (a) 及び (b) を比較すると，CIDEr-D はそれぞれ 0.06 及び 1.57 ポイント低下した．これより，transformer デコーダが最も性能向上に寄与していることがわかった．

同様に，BILA-caption データセットにおいて，提案手法と条件 (a) 及び (b) を比較すると，CIDEr-D はそれぞれ 4.96 及び 4.56 ポイント低下した．これより，RSA 層が最も性能向上に寄与していることがわかった．

5. おわりに

本稿では，将来イベントについての説明文を生成するタスクの一つである，Future captioning タスクを扱った．特に，日常生活における Future captioning タスクを扱った．

本稿の主要な貢献は以下である．

- 将来イベントについての説明文を生成可能なクロスモーダル言語生成モデル， RFCM を提案した．
- RSA エンコーダの導入により，効果的にイベント間の関係性を抽出できる．
- RFCM は YouCook2-FC 及び BILA-caption データセットにおいてベースライン手法を上回った．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST CREST，NEDO の助成を受けて実施されたものである．

参考文献

[Banerjee 05] Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in *the ACL Workshop on IEEM for MTS*, pp. 65–72 (2005)

[Devlin 19] Devlin, J., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *NAACL-HLT*, pp. 4171–4186 (2019)

[Ging 20] Ging, S., et al.: COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning, in *NeurIPS*, pp. 22605–22618 (2020)

[Hosseinzadeh 21] Hosseinzadeh, M., et al.: Video Captioning of Future Frames, in *WACV*, pp. 980–989 (2021)

[Kambara 21] Kambara, M. and Sugiura, K.: Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions, *IEEE RA-L*, Vol. 6, No. 4, pp. 8371–8378 (2021)

[Kim 21] Kim, M., Kwon, H., Wang, C., et al.: Relational Self-Attention: What’s Missing in Attention for Video Understanding, in *NeurIPS* (2021)

[Krishna 17] Krishna, R., Hata, K., et al.: Dense-Captioning Events in Videos, in *CVPR*, pp. 706–715 (2017)

[Lei 20] Lei, J., Wang, L., et al.: MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, in *ACL*, pp. 2603–2614 (2020)

[Lin 04] Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries, in *Text summarization branches out*, pp. 74–81 (2004)

[Magassouba 21] Magassouba, A., Sugiura, K., et al.: Predicting and attending to damaging collisions for placing everyday objects in photo-realistic simulations, *Advanced Robotics*, Vol. 35, No. 12, pp. 1–13 (2021)

[Mahmud 21] Mahmud, T., Billah, M., et al.: Prediction and Description of Near-Future Activities in Video, *CVIU*, Vol. 210, p. 103230 (2021)

[Miech 19] Miech, A., et al.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in *ICCV*, pp. 2630–2640 (2019)

[Miech 20] Miech, A., Alayrac, J.-B., et al.: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, in *CVPR*, pp. 9879–9889 (2020)

[Mori 21] Mori, Y., Hirakawa, T., et al.: Image Captioning in Near Future from Vehicle Camera Images and Motion Information, in *IEEE IV*, pp. 1378–1384 (2021)

[Prechelt 98] Prechelt, L.: Automatic early stopping using cross validation: quantifying the criteria, *Neural Networks*, Vol. 11, No. 4, pp. 761–767 (1998)

[Vaswani 17] Vaswani, A., Shazeer, N., et al.: Attention Is All You Need, in *NeurIPS*, pp. 5998–6008 (2017)

[Vedantam 15] Vedantam, R., et al.: CIDEr: Consensus-based Image Description Evaluation, in *CVPR*, pp. 4566–4575 (2015)

[Wang 18] Wang, X., et al.: Video Captioning via Hierarchical Reinforcement Learning, in *CVPR*, pp. 4213–4222 (2018)

[Zhou 18] Zhou, L., Xu, C., and Corso, J.: Towards Automatic Learning of Procedures From Web Instructional Videos, in *AAAI*, pp. 7590–7598 (2018)